基于依存关系的中文微博作者性别识别*

祁瑞华
(大连外国语大学软件学院　大连　116044)

摘要: 【**目的**】针对网络文本篇幅短小、传统文体特征集稀疏等特点, 探讨依存关系在中文微博作者性别识别中的应用。【**方法**】选取腾讯公开微博作为实验语料, 抽取依存关系特征与现有文献中的词汇特征、结构特征、功能词特征、词性标注特征和微博特征进行对照实验。【**结果**】采用支持向量机、朴素贝叶斯、最近邻和决策树算法的对照实验验证了本文方法在中文微博作者性别识别任务中的准确率、召回率和 F-Measure 最高。【**局限**】依存关系在微博作者性别识别中的有效性还需在大规模语料上进一步验证。【**结论**】本文模型能够避免短文本特征集的稀疏性, 与其他对照特征集相比, 能更有效地识别作者性别。

关键词: 依存关系　中文微博　性别识别
分类号: TP182

1 引　言

网络文本随着各种网络应用的快速普及而大量涌现, 作者身份属性分析在市场营销、网络取证等领域的应用已经成为热点。Twitter 平台上每天新增的信息在 5 亿条以上, 而与此同时用户身份频频被盗用, 仅 2016 年就有超过 3 200 万 Twitter 用户的登录信息被泄露[1], 此后的 Twitter 身份盗用案例逐年增加。网络社交媒体用户量和信息量的激增进一步凸显了作者身份属性研究的迫切性。

作者性别分析是身份属性研究的主要任务之一, 网络文本作者性别分析有助于商家针对客户群体开展精准营销, 从而提高个性化推荐和拓展市场的效率。作者的性别分析还有助于鉴别匿名虚假信息和不实言论的来源, 避免对社会经济秩序和治安造成严重负面影响。

微博已成为作者性别分析关注的重要领域, 2016 年第一季度仅新浪微博平台的月活跃用户数同比增长 32%, 已达到 2.61 亿[2]。微博作者的性别识别已成为国内外研究的热点, 例如利用 Twitter 用户信息和 Tweets 的内容判断作者性别[3], 或是利用中文微博用户名和微博文本构建作者性别分类融合器[4]等。现有方法的局限在于对用户名等信息的依赖, 未考虑作者刻意隐藏身份的情况。

为此, 本文提出无需微博用户信息的作者性别识别方法, 通过抽取微博文本的依存关系特征构建微博作者性别文体特征模型, 并在微博语料上与现有文献中的特征集进行比较, 验证依存关系特征在微博作者性别识别中的有效性。

2 作者性别识别相关研究

网络文本作者性别分析研究涉及网络评论、BBS 和博客等语料, 以英文为主。代表研究有 Schler 等[5]分析了数万篇近 3 亿单词的英文博客语料, 证实了男性与女性在写作风格和内容方面均存在明显区别。Argamon 等[6-7]结合人称代词、限定词、介词、内容特征等语言学特征和 Bayesian Multinomial Regression 算法对博客作者语料进行作者性别分析, 实验结果达到 70%左右的准确率。此外, 在希腊文语料上, Mikros 等[8]利用 20 位作者的博客语料, 建立包括词长统计、

通讯作者: 祁瑞华, ORCID: 0000-0002-2583-3055, E-mail: rhqi@dlufl.edu.cn。

*本文系国家社会科学基金一般项目“典籍英译国外读者网上评论观点挖掘研究”(项目编号: 15BYY028)和国家教育部回国人员科研启动基金项目(项目编号: 教外司[2015]1098)的研究成果之一。

词汇丰富度、最常用词汇和字符 Ngram 等特征的文体特征集，采用支持向量机算法得到 80%以上的性别识别准确率。Rangel 等[9]提出词频、标点、词性标注、英文和西班牙语情感词等文体特征有助于鉴定匿名作者的性别，并采用支持向量机算法在 PAN-AP-133 数据集上取得 57%的性别识别正确率。上述研究中的文本长度普遍高于微博文本，特征集从数百维到数千维，作者特征集存在明显的稀疏性。此外，Burger 等[3]抽取 Twitter 用户的昵称、账户名、个人描述和 Tweets 内容的字符 1-5gram 和单词 1-2gram 判断作者的性别，得到了最高 92%的准确率，只采用 Tweets 文本特征时，仅取得 75%左右的作者性别识别准确率。

针对中文语料，唐琴等[10]提取中文小说中的性别倾向描述词和称谓词，指出前者具有更好的性别指示作用，并利用特征合集在人名性别识别实验中取得 73.2%的正确率，此方法未在短文本语料上验证。黄发良等[11]基于词项特征向量模型提出粗糙集微博用户性别识别算法，其改进的特征词频数加权机制降低了文档零相似现象，但未提出如何确定容差阈值。白丽娟[12]选取天涯网站汽车和股票论坛文本，通过 CFS 和 BestFirst 算法得到特征词，采用朴素贝叶斯和支持向量机等算法获得 70%-80%的准确率，此方法准确率依赖于文本长度，基于内容的特征词虽然能够提高性别识别的准确率，但影响了方法的跨主题适用性。王晶晶等[4]在中文微博上采用用户名 1-2gram 和首位字特征，与微博文本的 1-2gram 特征构建贝叶斯分类融合算法，达到最高 90%左右的作者性别识别准确率，但只采用微博文本特征的最高准确率仅为 74%左右。

深层句法依存关系分析能够提取主题无关的抽象句法结构信息，从而发现隐含的写作习惯[13]，近年来被尝试应用于作者风格分析，例如 Hollingsworth[14]采用 DepWords 编码替代传统句法依存关系，并利用其统计特征识别英文侦探小说的作者身份；Zhang 等[15]抽取包括结构特征、功能词、POS、常用词和依存关系等特征，在 21 本英文作品和路透社语料上的对照实验表明依存关系有助于提高作者身份识别效率。依存关系在作者性别识别中的效果还有待探索。

本文基于对现有研究和微博文本特征的分析，提出新的基于依存关系的作者性别文体特征模型。

3　作者性别文体特征模型

设作者性别集合为 $A=\{Female, Male\}$，有训练样本集 $T_i=\{t_{i1}, t_{i2}, \cdots, t_{ij}\}$，作者性别自动识别的任务是学习训练集建立作者性别特征模型，并根据此模型为匿名文本 t 指定一个最可能的作者性别 G_t ($G_t \in A$)。为完成这一任务，首先要将非结构化的文本映射到文体特征向量空间并抽取作者性别文体特征集，此文体特征集应该具有区分作者性别的描述能力，其中的特征值应具有较好的可获取性。

3.1　依存关系

依存关系是由法国语言学家 Tesnière 等[16]提出描述句法结构的理论框架，其描述的基础是词与词之间的从属和支配关系，目前已经广泛应用于文本挖掘、多语言处理、语义标注和信息检索等领域。依存关系由句子核心词和依存词的依存关系对组成，句子 $S=\{w_0, w_1, \cdots, w_n\}$ 中，w_i 为句子中第 i 个词，抽取句子的依存关系后，句子可表示为 $S=\{R_1(w_{11}, w_{12}), R_2(w_{21}, w_{22}), \cdots, R_m(w_{m1}, w_{m2})\}$，其中每个依存关系 R_i 是由 (w_{i1}, w_{i2}) 词对构成的有向弧，由支配词 w_{i2} 指向被支配词 w_{i1}，其中 $w_{ij} \in S$，$r_1, r_2, \cdots, r_m \in R$，$R$ 为所有依存关系类型的集合。依存关系的形式化描述公理包括[17]：一个句子的独立成分只有一个；除此独立成分，句子中其他成分都直接依存于本句中的某一成分；句中任何成分不能依存于两个以上成分；若成分 X 直接依存于成分 Y，成分 Z 在句子中的位置位于 X 和 Y 之间，则成分 Z 依存于成分 X 或 Y，或依存于 X 和 Y 之间的某一成分。

依存关系作为作者性别文体特征具有三个优势：依存关系存储结构简单，可计算性好，对网络文本大数据和跨语言环境具有良好的适应性；依存句法分析强调句子成分间的支配与被支配、修饰与被修饰的依存关系，不限于句子成分顺序的特性有助于分析句式变化灵活的网络文本；此外，依存关系提取抽象句法结构信息，具有内容无关性。

本文以复旦大学FudanNLP[18]中定义的 22 种汉语依存关系作为作者性别识别特征集，$F_{依存关系}$ ={关联，主语，标点，疑问连动，补语，语态，的字结构，介宾，数量，宾语，地字结构，感叹，时态，之字结构，同位语，得字结构，并列，连动，修饰，核心词，定语，状语}，句子成分依存关系如图 1 所示。

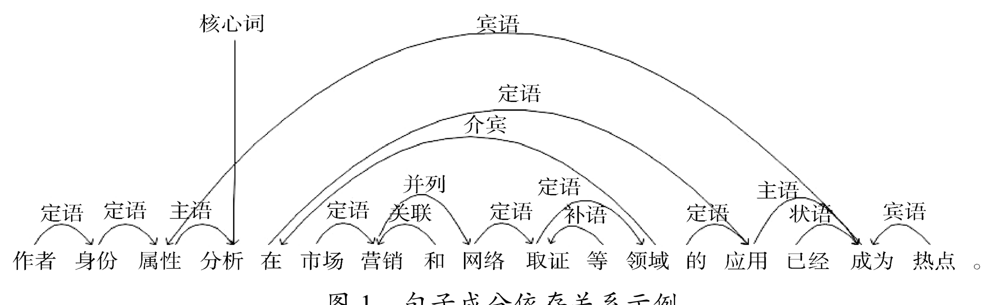


图 1 句子成分依存关系示例

3.2 现有文体特征

本文在对照实验中引入现有文献提出的主要文体特征，包括词汇特征、结构特征、功能词特征、词性标注特征和微博特征。其中，词汇特征包括单词的统计特征和频率，如词长、词汇丰富度、词频、单词Ngram以及特殊词汇等，词汇特征的抽取很大程度上依赖于语料长度，因此通常不单独使用。考虑到微博篇幅短小，为避免词汇特征稀疏，本文根据国家语言资源监测与研究中心发布的2015年中国语言生活状况绿皮书[19]，在对照实验中选取内容无关的数词、高频词、时间词和日期词的出现次数作为词汇特征。

结构特征包括文本组织和布局相关的特征，包括标点符号、段落数、段落长、平均句长等，在E-mail、博客或微博等短文本上尤为有效。本文根据文献[15]在对照实验中选取句子个数、字符数以及冒号、分号、千百分号、单位符号、句号、左右引号、左右括号、逗号、叹号、省略号(单)、省略号(双)、破折号、空格、问号和顿号出现的次数作为结构特征。

功能词特征指本身并没有独立完整词汇意义，只表达语法意义或语法功能的词，具有与主题内容无关的特点。现代汉语中的功能词也称为虚词，功能词出现频率高、数量少，已经被证实是有效的文体风格特征[20]。中文功能词担负着西方语种中实词变化表达的语法意义，具有更重要的语法作用。在对照实验中选取文献[20-21]中的中文功能词的合集作为功能词特征。

词性标注特征是根据词形或句法行为作用进行的单词类型标注，通常不涉及词的具体含义，具有主题无关性。中国科学院计算技术研究所的ICTCLAS[22]汉语词性标注集包括22个一类标记，66个二类标记和11个三类标记，本文统计ICTCLAS汉语词性标注集22个一类词性POS标注在每千词中出现的次数。

微博特征包括微博文本特有的文本布局格式，例如话题的引用、用户名的指代、图片超链接的使用等。本文参照文献[23]统计微博中图片出现的次数、网址URL、#符号、@符号、Email和表情符号出现的频次。

4 微博作者性别识别实验

4.1 数据准备

选取腾讯公开微博作为实验语料，收集腾讯微博实名注册热点人物在2012年10个月期间的微博文本合计6 530篇，其中男性作者5 496篇，女性作者1 034篇。语料中最长微博文本篇幅为284字符，最短微博文本为5字符，平均文本长度为73字符，语料中100字符以下的样本占65%。

实验采用ICTCLAS 2015[22]进行中文语料分词和词性标注，采用FudanNLP1.5[18]分析依存句法关系，分类算法实验环境为Weka 3.7.9[24]。在对照实验中执行十折交叉验证，以作者性别识别的准确率(Precision)、召回率(Recall)和F-Measure评估模型的性能。

将汉语依存关系特征集与文献[20-21, 23]中的主要文体特征进行对比实验，主要包括：词汇特征、结构特征、功能词特征、词性标注特征和微博特征，特征集及其维度和关系如图2所示。

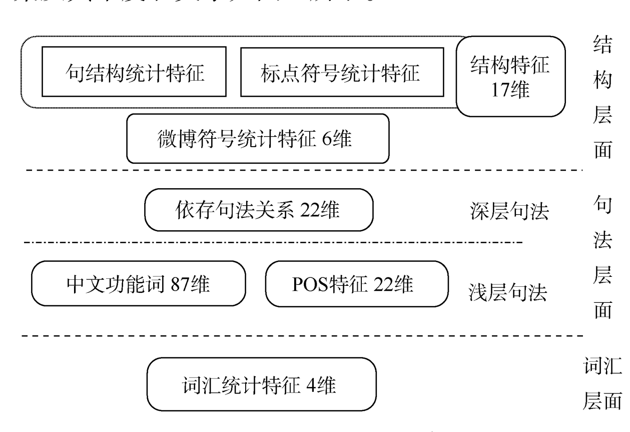


图 2 对照实验采用的特征集

4.2　实验及分析

为验证依存关系在中文微博作者性别识别中的有效性，采用支持向量机(LibSVM)、朴素贝叶斯(NBC)、最近邻(IBK)和决策树(C4.5) 4 种分类算法进行对比实验。实验结果如表 1 所示，4 种算法作者性别识别的最高值已经加粗显示。

表 1　LibSVM、NBC、IBK 和 C4.5 中文微博作者性别识别结果

算法	指标	词汇特征	结构特征	微博特征	功能词	词性标注	依存关系
Lib-SVM	Precision	0.797	0.897	0.918	0.832	0.861	**0.998**
	Recall	0.843	0.903	0.921	0.852	0.868	**0.998**
	F-Measure	0.787	0.898	0.914	0.802	0.835	**0.998**
NBC	Precision	**0.838**	0.799	0.766	0.828	0.798	0.814
	Recall	0.396	0.815	0.806	0.436	**0.834**	0.691
	F-Measure	0.432	0.806	0.781	0.482	**0.807**	0.730
IBK	Precision	0.809	0.912	0.909	0.806	0.834	**0.999**
	Recall	0.811	0.913	0.914	0.812	0.836	**0.999**
	F-Measure	0.810	0.912	0.909	0.809	0.835	**0.999**
C4.5	Precision	0.824	0.928	0.918	0.899	0.851	**0.997**
	Recall	0.852	0.929	0.921	0.904	0.864	**0.997**
	F-Measure	0.818	0.928	0.915	0.893	0.855	**0.997**

(1) 比较各特征集对作者性别的区分效果，总体上依存关系特征集在中文微博数据集实验中的准确率、召回率和 F-Measure 值最高，在支持向量机、最近邻和决策树算法的实验中三个关键指标值均达到 99.7%以上。实验结果证实了依存关系特征集能够挖掘不同性别作者在微博文本表达中的深层句法特征，与词汇、结构、功能词、词性和微博特征比较更适应短文本，能够避免特征集稀疏对算法效率的影响。

(2) 从算法性能看，总体上，最近邻、支持向量机和决策树 C4.5 算法的作者性别识别准确率、召回率和 F-Measure 的加权平均值较高，朴素贝叶斯算法的效果一般，分析其原因是朴素贝叶斯算法的独立性假设在大多数特征集中并不成立，而最近邻算法和支持向量机算法能够适应样本中的噪音，决策树算法采用信息增益率作为特征选择依据，克服了短文本的特征稀疏和噪声干扰。

(3) 在朴素贝叶斯算法作者性别识别实验中，依存关系特征集的效果不如其他特征集，分析其原因是依存关系特征不满足朴素贝叶斯算法的独立性假设。

(4) 从分析关键特征的角度看，图 3 是决策树 C4.5 算法对微博文本依存关系特征集进行性别识别时构造的决策树，决策树中关键特征比较集中，包括关联关系、主语关系、时态关系、之字结构、得字结构和连动关系，可以作为主要特征进一步探究不同性别作者在句法结构选用中的倾向。

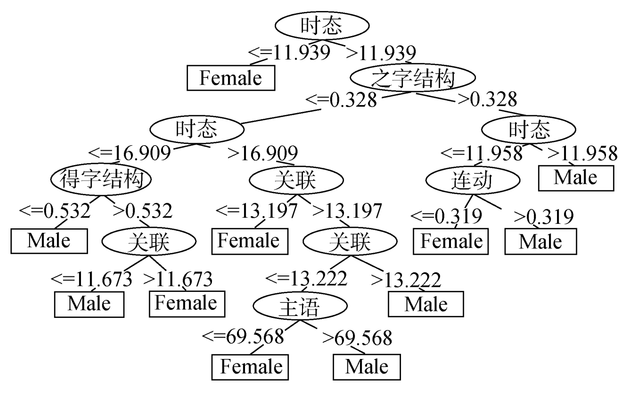


图 3　C4.5 算法依存关系特征集决策树

5　结　语

本文探究了深层句法分析特征在中文微博作者性别分析中的应用。实验结果表明，与现有文献中的方法相比，本文提出的基于依存关系的中文微博作者性别文体特征模型能够避免短文本特征集的稀疏性，与其他对照特征集相比，能更有效地识别作者性别。本文发现依存关系特征中的关联关系、主语关系、时态关系、之字结构、得字结构和连动关系在决策树中起到了关键节点的作用，后续研究将在大规模语料上进一步验证。

参考文献：

[1] 新浪科技.3200 万 Twitter 账号被盗 [R/OL].[2016-06-09]. http://tech.sina.com.cn/i/2016-06-09/doc-ifxszmaa1783949.shtml. (Sina Science and Technology. 32 Million Twitter Account Stolen [R/OL]. [2016-06-09]. http://tech.sina.com.cn/i/ 2016-06-09/doc-ifxszmaa1783949.shtml.)

[2] 新浪科技.微博月活跃用户增至 2.61 亿[R/OL]. [2016-05-12]. http://tech.sina.com.cn/i/2016-05-12/doc-ifxsenvm0294013.shtml. (Sina Science and Technology. Micro-blog Monthly Active Users Increased to 261 Million[R/OL].[2016-05-12]. http://tech.sina.com.cn/i/2016-05-12/doc-ifxsenvm0294013.shtml.)

[3] Burger J D, Henderson J, Kim G, et al. Discriminating

Gender on Twitter[C]//Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011: 1301-1309.

[4] 王晶晶, 李寿山, 黄磊.中文微博用户性别分类方法研究[J]. 中文信息学报, 2014, 28(6): 150-155, 168. (Wang Jingjing, Li Shoushan, Huang Lei. User Gender Classification in Chinese Microblog [J]. Journal of Chinese Information Processing, 2014, 28(6): 150-155, 168.)

[5] Schler J, Koppel M, Argamon S, et al. Effects of Age and Gender on Blogging [C]// Proceedings of the 2006 Association for the Advance of Artificial Intelligence Spring Symposium: Computational Approaches to Analyzing Weblogs. 2006.

[6] Argamon S, Koppel M, Pennebaker J W, et al. Automatically Profiling the Author of an Anonymous Text [J]. Communications of the ACM, 2009, 52(2): 119-123.

[7] Argamon S, Koppel M. A Systemic Functional Approach to Automated Authorship Analysis [J]. Journal of Law & Policy, 2013, 12: 299-315.

[8] Mikros G K, Perifanos K. Authorship Attribution in Greek Tweets Using Author's Multilevel N-Gram Profiles[C]// Proceedings of the 2013 Association for the Advance of Artificial Intelligence (AAAI) Spring Symposium: Analyzing Microtext. 2013.

[9] Rangel F, Rosso P. Use of Language and Author Profiling: Identification of Gender and Age[C]//Proceedings of the 10th Workshop on Natural Language Processing and Cognitive Science. 2013.

[10] 唐琴, 林鸿飞. 文本中人物性别识别研究[J]. 中文信息学报, 2010, 24(2): 46-51. (Tang Qin, Lin Hongfei. Research on Gender Recognition for Character in Text [J]. Journal of Chinese Information Processing, 2010, 24(2): 46-51.)

[11] 黄发良, 熊金波, 黄添强, 等. 基于粗糙集的微博用户性别识别[J]. 计算机应用, 2014, 34(8): 2209-2211. (Huang Faliang, Xiong Jinbo, Huang Tianqiang, et al. Gender Identification of Microblog Users Based on Rough Set[J]. Journal of Computer Applications, 2014, 34(8): 2209-2211.)

[12] 白丽娟. 基于文本挖掘的性别分类研究[D]. 哈尔滨: 哈尔滨工业大学, 2011. (Bai Lijuan.Gender Classification Based on Text Mining [D]. Harbin : Harbin Institute of Technology, 2011.)

[13] 祁瑞华, 杨德礼, 郭旭, 等. 基于多层面文体特征的博客作者身份识别研究[J]. 情报学报, 2015, 34(6): 628-634. (Qi Ruihua, Yang Deli, Guo Xu, et al. Blogger Identification Based on Multidimensional Stylistic Features[J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(6): 628-634.)

[14] Hollingsworth C. Using Dependency-based Annotations for Authorship Identification [M]. Text, Speech and Dialogue, Springer Berlin Heidelberg, 2012: 314-319.

[15] Zhang C, Wu X, Niu Z, et al. Authorship Identification from Unstructured Texts[J]. Knowledge-Based Systems, 2014, 66: 99-111.

[16] Tesnière L, Osborne T, Kahane S. Elements of Structural Syntax[M]. John Benjamins Publishing Company, 2015.

[17] Robinson J J. Dependency Structures and Transformational Rules[J]. Language, 1970, 46(2): 259-285.

[18] Fudan Natural Language Processing Group. FudanNLP [EB/OL]. [2016-01-01].http: //nlp.fudan.edu.cn/ software/.

[19] 国家语言资源监测与研究中心平面语言媒体中心. 历年中国语言生活状况绿皮书[R/OL]. [2015-01-01]. http://cnlr.blcu.edu.cn/col/col8765/index.html. (National Language Resources Monitoring and Research Center. Chinese Language Situation over the Years [R/OL]. [2015-01-01]. http://cnlr.blcu.edu.cn/col/ col8765/index.html.)

[20] Zheng R, Li J, Chen H, et al. A Framework for Authorship Identification of Online Messages: Writing-style Features and Classification Techniques [J]. Journal of the American Society for Information Science and Technology, 2006, 57(3): 378-393.

[21] Yu B. Function Words for Chinese Authorship Attribution [C]// Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2012.

[22] ICTCLAS 2015 [EB/OL]. [2015-01-01]. http://ictclas.nlpir.org/downloads.

[23] Silva R S, Laboreiro G, Sarmento L, et al. 'twazn me!!!; (' Automatic Authorship Analysis of Micro-blogging Messages [M]. Natural Language Processing and Information Systems. Berlin Heidelberg: Springer, 2011: 161-168.

[24] Machine Learning Group at the University of Waikato. WEKA [EB/OL]. [2015-01-01]. http://www.cs.waikato.ac.nz/ml/weka/downloading.html.

利益冲突声明:

作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: rhqi@dlufl.edu.cn。

[1] 祁瑞华. Chinese-microblog.txt. 腾讯微博原始抓取数据.
[2] 祁瑞华. female-male.csv. 腾讯微博标注特征集.

收稿日期: 2016-10-06
收修改稿日期: 2016-11-29

Identifying Chinese Microblog Author Gender Based on Dependency

Qi Ruihua
(School of Software, Dalian University of Foreign Languages, Dalian 116044, China)

Abstract: [**Objective**] This paper proposes a new method to indentify the gender of Chinese microblog author with the help of dependency features. [**Methods**] This study collected public posts from Tencent Microblogs and extracted the dependency features, which were analyzed and compared with existing vocabulary, structure, function words, and part-of-speech tagging features. [**Results**] A controlled experiment showed that the proposed method obtained the highest values of precision, recall and F-measure. [**Limitations**] The new method needs to be examined with larger corpus. [**Conclusions**] The proposed method is the most effective way to identify the gender of microblog author.

Keywords: Dependency　Chinese Microblog　Gender Identification

ProQuest 发布白皮书《机遇和挑战：电子书，印刷本，选择对图书馆和读者的影响》

ProQuest 于近日发布了题为《机遇和挑战: 电子书, 印刷本, 选择对图书馆和读者的影响》的白皮书, 现已可供下载。该书专注于英国高等教育图书市场, 将来自英国图书馆员的评论与全球数据结合起来, 分析了管理图书馆馆藏的复杂性以及不断变化的内容类型给整个业界带来的机会。

受英国学术图书馆馆员讨论的启发, 该书深入研究了 ProQuest 所开发的一系列使得图书馆管理印刷和数字内容工作流程更加高效、灵活的解决方案。除此之外, 该书还提供了对当前图书大环境及其复杂性的概述。该书重点研究了图书馆如何在逐渐流行的既便捷又可访问的电子书与持续的印刷本需求之间进行平衡。

ProQuest 图书部高级副总裁兼总经理 Kevin Sayar 表示: “图书馆员的具体目标不尽相同, 但所有图书馆都致力于将研究人员与他们需要的内容联系起来, 并在有限的预算内策划尽量丰富、均衡的馆藏资源。我们期待着该白皮书能激发起我们所有人就建立一个真正支持研究和学习的全球性的图书大环境这一主题进行对话。”

《机遇和挑战: 电子书, 印刷本, 选择对图书馆和读者的影响》可以从 http://www.proquest.com/documents/Whitepaper-Obstacles-and-Opportunities.html 下载。

(编译自: http://www.proquest.com/about/news/2017/Explores-Obstacles-and-Opportunities-in-Managing-Book-Collections.html)

(本刊讯)